Stock Price Prediction with Time Series Forecasting

Vedant Singhvi, *Student, CSYE6530*, Suraj Ghule, *Student, CSYE6530*

Abstract—Stock price prediction is the process of trying to predict the future prices of a stock or any other financial instrument traded on an exchange. Stock price prediction has been a strong topic of discussion with many trying to predict prices to yield profits. Initially a far-fetched dream, data science has helped make huge strides in stock price prediction.

The purpose of the paper is to analyze and understand the advantages and disadvantages of various algorithms applicable on Time Series Data. The dataset used is the Standard and Poor's 500 Dataset, with special focus on Apple stock price. After the elimination of stationarity, trend and seasonality, the closing price of the Apple Stock was predicted using multiple algorithms which included:

- Moving Average Model
- Autoregression Model
- ARIMA
- Facebook's Prophet
- Recurrent Neural Networks
- Recurrent Neural Network with Long Short-Term Memory

Index Terms—Stock Price, Time series, OHLC, candlestick chart, trend, seasonality, stationarity and non- stationarity time series, Dickey Fuller test, Residual, Auto-Regression, Moving Average, p-value, ARIMA, Facebook Prophet, LSTM, Artificial Neural Network, RNN, feed-forward neural network, RMSE, exponential smoothing, log-difference.

Vedant Singhvi
Singhvi.v@husky.neu.edu

Suraj Ghule
Ghule.s@husky.neu.edu

Information Systems Program
Northeastern University, Boston, MA, USA

I. INTRODUCTION

Time series is a sequence of well-defined data points measured at a consistent time interval over a period. A time series can be anything from a person's heart rate to daily closing price of a company stock. Basically, time series is the data which is dependent on time. Stock price time series analysis is a complicated and a controversial topic to say the least. Stock price of a company depends on a plethora of factors which makes the prediction a challenging task for mathematicians and financial gurus. The advent of Machine Learning and eventually Deep Learning proved out to be a massive push in this field resulting in companies hiring Data Scientists for Time Series Forecast and Analysis.

Generally, the stock price data shows extensive variations owing to several factors like the government, financial situation of the company, actions of the competitors, etc. The challenge in hand of a Data Scientist is to take in account all the possible factors and predict the best possible price of the stock. Nowadays, the prediction model/system are advances enough that they help the Data Scientist by themselves making well informed choices and providing quick results which would generally require extensive calculations.

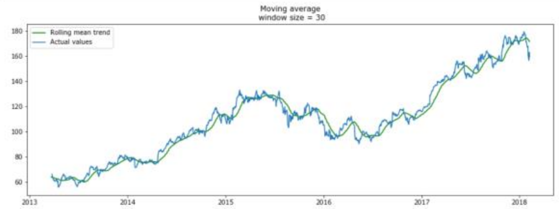
Stock price prediction using Machine Learning depends on several factors including Trend, Seasonality, Stationarity and Autocorrelation. A good Machine Learning model would take care of each of the factor and then predict the price. This would result in the best possible prediction. Several methods are available to analyze and take care of them.

II. Data Acquisition

In this paper and project, the S&P 500 (Standard & Poor's) dataset is used, which consists of the stock prices of top 500 companies which includes the High, Low, Opening and Closing Price from the years 2013 to 2018. This paper primarily focuses on the Apple Inc stock price, with emphasis only on the Closing Price.

III. Data Exploration

Data Exploration involved comparing the Closing Price against time. Changing the time windows provided extra insight into the movement of Apple's Closing price. In addition to that, the time series graph was 'smoothened' and the noise was filtered by calculating the Moving Average of the prices in multiple time windows. Moving Average is calculated by taking the arithmetic mean of a given set of values and is used to gauge the directions of the current trends. Also, OHLC and Candlestick charts provided additional information.



IV. Data Preprocessing

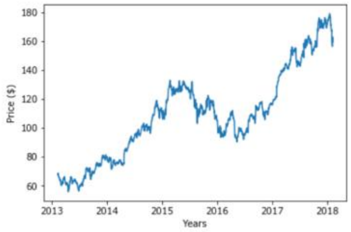
Data Preprocessing in Time Series can be split into 2 parts:

1. Trend, Seasonality and Stationarity Detection
2. Trend, Seasonality and Stationarity Elimination

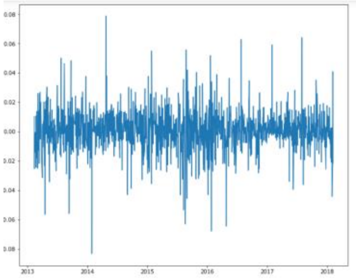
Time series can pose several problems to Data Scientist. Since time series is dependent on 'time', it can exhibit trend and seasonality, which in turn can make a time series non-stationary and have high auto correlation which is undesirable.

In this project, inconsistencies in the time series were detected by applying Dickey Fuller test to check the Stationarity, decomposing the time series in Residual, Trend and Seasonality and plotting ACF and PACF charts.

A time series can be made stationary by applying Differencing and Transformation method. In this work, applying logarithmic function and taking the difference between the data points made the time series stationary and resulted in the p value reduce from 0.87 to 0.00.



Non-Stationary



Stationary

V. Models Applied

Auto-Regression Model: Auto Regression model uses the dependent relationship between an observation and some number of lagged observations. It is often considered as one of the most basic time series algorithms. In a nutshell, it a linear regression model applied against the previous data points and values. It takes in a single parameter, which is 'p'. 'p' basically is the number of autoregressive terms for the model. It can be estimating by understanding the ACF graph. For this work, a p value of 3 was considered suitable.

Moving Average Model: It is a model that uses the dependency between an observation and residual errors from a moving average model applied to lagged observations. It is similar Auto-Regressive Model. It takes in a single parameter, which is 'q'. 'q' is the number of nonseasonal differences. It can be estimating by understanding the PACF graph. For this work, a q value of 3 was considered suitable.

ARIMA: ARIMA, or Auto-Regressive Integrated Moving Average models are statistical models specifically used for time series analysis and forecasting. It is widely used and is expected to provide robust predictions. Also referred to as Box Jenkins methodology, it consists of 3 steps:

1. Identification: Used to summarize and sub-class the data
2. Estimation: Trains the parameters of the model
3. Diagnostic Checking: Evaluates the model and checks for area of improvement

The parameters that ARIMA takes in are p, q and d. 'p' and 'q' are the same as the Autoregression model and Moving Average model, with 'd' being the degree of differencing.

Facebook Prophet: Facebook Prophet is an open source software/library released by Facebook's Data Science team, specially created for time series analysis and forecasting. It is extremely easy to implement, with the major processing being taken care by Prophet itself.

The major features for Facebook Prophet include highly accurate and fast, fully automatic and tunable forecast.

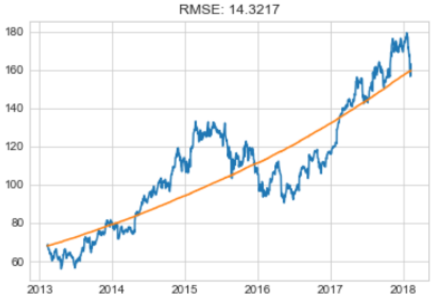
Since the software is mostly automated, minimal data-preprocessing was required.

LSTM: LSTM is a type of Artificial Neural Network architecture in the field of deep learning. It has a feedback connection which make it a general-purpose computer, which means that it can be used anywhere and makes it a very powerful model since it can not only process single data points, but also entire sequences of data.

RNN: It is a class of Neural Networks wherein the connection between nodes form a directed graph along a temporal sequence which allow them to have temporal behavior. They have an advantage over feed-forward neural networks which allows them to use their internal memory to process sequence of inputs.

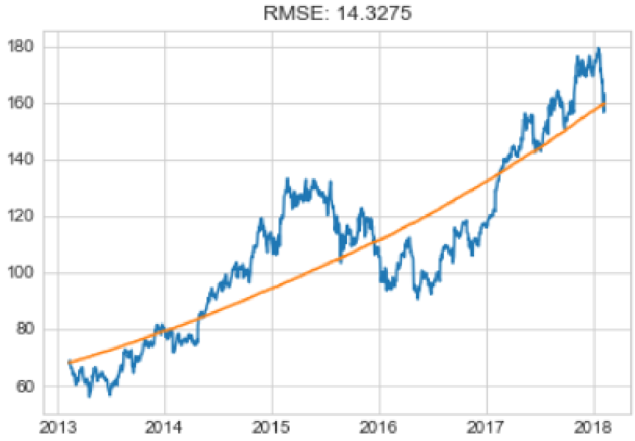
VI. RESULTS

1. Results using Auto Regressive Model:



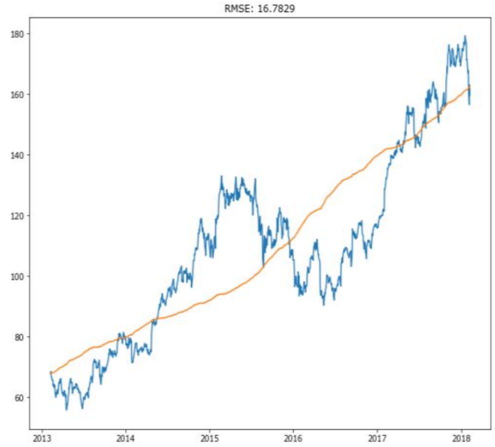
The RMSE value using AR Model is 14.3217.
Advantages of AR model is that it is easy to implement and requires minimal parameters.
Disadvantages include lack of complexity and not so accurate predictions.

2. Results using Moving Average Model:



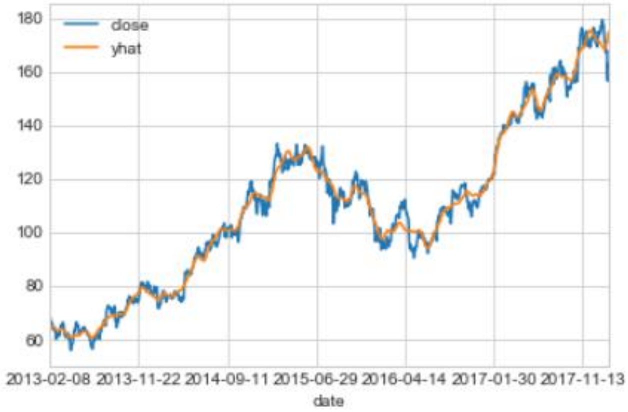
The RMSE value using MA Model is 14.3275.
The predicted values show similar trend with actual values but are again not much nearer to actual values.
Advantages of MA model is that it is easy to implement and requires minimal parameters.
Disadvantages include lack of complexity and not so accurate predictions.

3. Results using ARIMA Model:



The RMSE value using ARIMA Model is 16.78.
Although more than the basic models, it provides good variations and deviations in the data.
Advantages of ARIMA model is that it is easy to implement, requires minimal parameters and provides fairly accurate predictions.
Disadvantages include lack of complexity.

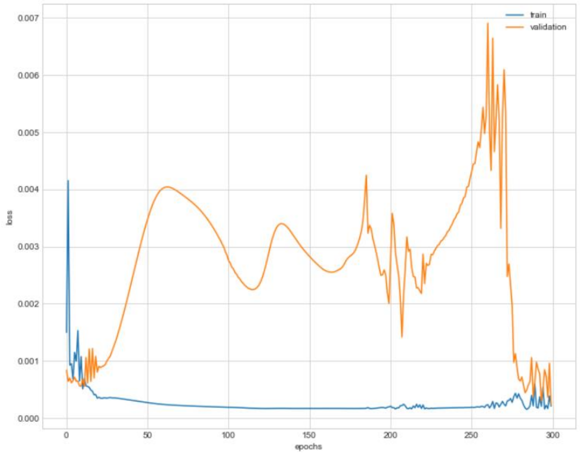
4. Results using Facebook prophet



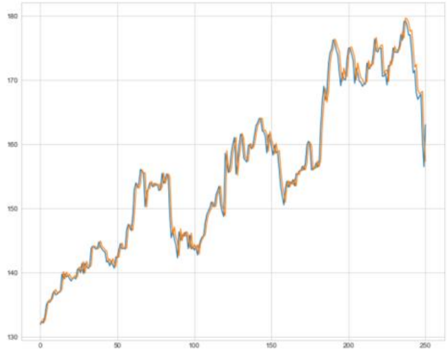
The RMSE value using Prophet is 3.23 which is better than AR, MA and ARIMA model. Additionally, the predicted values showed approximately same behavior as the actual values.
The advantages include fast computation, minimal tuning and accurate predictions.
Disadvantages include lack of support for multiplicative model.

5. Results using LSTM Model

a. Training and validation loss



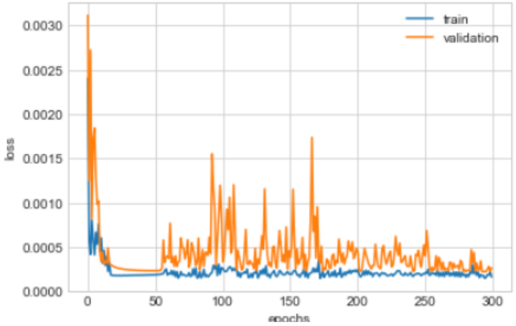
b. Predicted and actual values



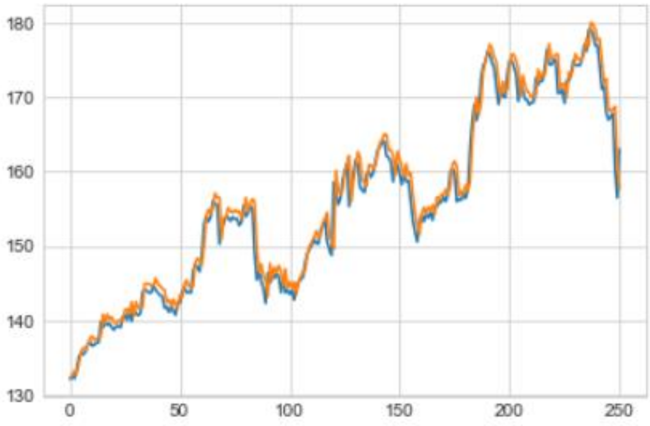
RMSE Value using LSTM is 12.426
The advantages include power of Neural Networks to exploit and use of prior data which is essential in Time Series.
Disadvantages include extensive parameter tuning and slow computation.

6. Results using RNN Model:

a. Training and validation loss



b. Predicted and actual values



RMSE Value using LSTM is 12.48
The advantages include power of Neural Networks to exploit and use of prior data which is essential in Time Series.
Disadvantages include extensive parameter tuning and slow computation

VII. Conclusion

Predicting stock prices is not an easy task primarily because of the insane volatile nature of the market. In fact, if everyone could predict stock prices, they probably would be rich by now! However, that is not the case.

In real life, stock prices require extensive market knowledge, and surely some analytics can be helpful. That is what we have done here.

Stock Price Time Series Forecasting is just an additional technique for stock price prediction.
After applying several algorithms and comparing them against each other. The RMSE values we recorded were:

1. AR model: 14.3217
2. MA model: 14.3275
3. ARIMA: 16.78
4. Prophet: 3.23
5. LSTM: 12.42
6. RNN: 12.48

Here, AR, MA and ARIMA are some basic time series forecasting models, which gave us a good prediction. One thing to keep in mind with Time Series is that a low RMSE value doesn't convey the complete story.

Among AR, MA and ARIMA model, ARIMA shows much better variation than AR and MA.
Facebook's Prophet is a promising algorithm, and expectedly performs well since it has been primarily been built for Time Series analysis.

The Neural Net duo performed well. It feels that the models tend to overfit a bit, and further investigation is required.

VIII. ACKNOWLEDGEMENT

We would like to show our gratitude to professor Nik Brown, and all the Teaching Assistants for guiding and encouraging us throughout the project.

IX. REFERENCES

[1] analyticsvidhya.com, ' A comprehensive beginner’s guide to create a Time Series Forecast (with Codes in Python)', [Online]

[2] machinelearningmastery.com, ' How to Create an ARIMA Model for Time Series Forecasting in Python', [Online]

[3] en.wikipedia.org 'Dickey–Fuller test', [Online]

[4] en.wikipedia.org ' A comprehensive beginner’s guide to create a Time Series Forecast (with Codes in Python)', [Online]

[5] kaggle.com, ' Time Series Forecast with Prophet', [Online]

[6] wikipdeia.com, ' Long short-term memory', [Online]

[7] machinelearningmastery.com, ' Stacked Long Short-Term Memory Networks', [Online]

[8] kaggle.com, ' Predict Stock Prices Using LSTM', [Online]

[9] kaggle.com, ' Deep-Learning-in-Python', [Online]

[10] stackoverflow.com, ' Simple Recurrent Neural Network input shape', [Online]

[11] ' Recurrent neural network', [Online].

[12] kaggle.com, ' everything-you-can-do-with-a-time-series', [Online]